The Report Committee for Xiao Bai
Certifies that this is the approved version of the following report:

Longitudinal Multilevel Models Analyzing the Trends of Land Use Effects on Non-Driving Travel Choice

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Susan N. Beretvas

Margaret Myers

Longitudinal Multilevel Models Analyzing the Trends of Land Use Effects on Non-Driving Travel Choice

by

Xiao Bai, B.S.; M.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

December 2013

Abstract

Longitudinal Multilevel Models Analyzing the Trends of Land Use Effects on Non-Driving Travel Choice

Xiao Bai, M.S. Stat.

The University of Texas at Austin, 2013

Supervisor: Susan N. Beretvas

Land use and transportation researchers have conducted numerous studies about land use effects on travel mode choice, and probed for effective policies to reduce driving, since less driving and more non-driving are widely recognized as more sustainable travel behaviors to resolve many environmental, energy and social equity issues. However, most of the previous studies rely on methodologies developed by cross-sectional data; only limited attention is explicitly given to explore the statistical techniques for longitudinal design and analysis. Using the neighborhood-level land use and persona-level travel mode choice data of 1997 and 2006 in the city of Austin, this paper attempts to establish and compare three distinct modeling approaches to analyze the trends of land use effects on people's choice behavior of non-driving travel mode. The three modeling approaches are: a comparison approach with two cross-sectional multilevel Logit models using single-year data, a pooling approach by building one multilevel model with two-year data, and a longitudinal multilevel model. Empirical modeling results indicate that the longitudinal multilevel model is the most reasonable

model for analyzing the longitudinal and multilevel datasets, since it is capable of estimating both time-invariant and time-variant land use effects, and internalizes time-variant random effects. The other two approaches may have several shortcomings. For example, the comparison approach fails to distinguish the time-variant and time-invariant effects; while the pooling model may lead to underestimated standard errors and t-statistics, and thus overestimate the significance of variables.

Table of Contents

List of Tables

List of Figures

1 Introduction

The growing public concern on the environmental, energy and social equity issues related to transportation has motivated a search for effective policy strategies to reduce driving, enhance access to travel alternatives, and achieve sustainable transportation. Land use policies, for example, mixed use development and transit-oriented development (TOD), are gaining popularity as a mobility tool to modify travel. The rational is that, if the land use is planned, designed, and regulated from automobile-oriented to become friendly to walking, biking, and public transit, driving would decrease; so do emissions and other traffic related issues. The last two decades have seen voluminous studies reporting empirical evidence to support (or challenge) the role of improving land use in increasing non-driving travel behaviors (Badoe and Miller 2000; Crane and Crepeau 1998; Crane 1999; Ewing and Cervero 2001, 2010).

Existing studies have largely confirmed the correlation between land use characteristics and travel behaviors, the focuses of which cover various travel types from commuting to nonworking travel behaviors (Handy 1994, 1996; Boarnet & Sarmiento 1998; Bhat et al. 1999, Handy & Clifton 2001; Boarnet & Crane 2001; Bhat & Gossen 2003; Rajamani 2003; Matt & Arentz 2003; Limanond & Niemeier 2004; Chen & Chen 2009; Chatman 2009), as well as cross different travel modes from auto driving, walking, to public transit (Greenwald & Boarnet 2001; Greenwald 2003; Rajamani 2003; Limanond & Niemeier 2004, Cao et al. 2009).

However, most of these land use-travel studies have conceptualized the influence of land use on travel in a "reduced-form" approach (Boarnet, 2011), which regresses travel behavior directly on land use pattern by assuming that land use characteristics have direct effects on travel and are statistically independent from other factors such as socioeconomic variables. Empirical evidence is thus collected to estimate the coefficients of land use attributes after controlling for other factors. This type of direct regression approach may have two limitations. First, it does not account for the potential contextual effect of land use. Land use may interact with personal/household characteristics and transportation services to influence travel decisions jointly. The second limitation pertains to statistical accuracy. Variables representing land use attributes are mostly area-based whereas variables measuring travel behaviors are more and more individual-based. Individuals located in the same area are subject to the same set of land use attributes. Regular regression analyses of individual's travel decisions will violate the assumption of independence of the sample when the land use variables enter the regression equation. Standard errors and t-statistic of regression coefficients will consequently be underestimated and coefficient estimates will be inefficient. Therefore, a number of studies have attempted to explore the limitations by applying multilevel analysis (Ma and Goulias 1997; Goulias 1999, 2002; Bhat 2000; Snellen et al. 2002; Schwanen et al. 2004).

Another deficiency of existing literature is that most observations on land use-travel connection rely on modeling analyses of cross-sectional data. The cross-sectional model is appropriate to reflect the static picture for a particular occasion, but hardly to capture the dynamic process of influence and distinguish the variation between the stable component (time-invariant effect) and the component that is subject to change over time (time-variant effect), and therefore is less powerful for causal inference (Rindfleisch et al. 2008). Thus, a longitudinal design is needed to overcome the deficiency of cross-sectional design. A longitudinal approach is an observational research method that involves studying the same group of individuals repeatedly over an extended period of time. Longitudinal studies allow researchers to distinguish short-term and unstable phenomena from long-term and relatively stable trends, and thus provide a meaningful understanding on development and time-span issues.

This paper thus attempts to explore statistical methodologies which are appropriate to study the longitudinal land use-travel connection. There are mainly two types of longitudinal studies: an aggregate trends study and an individual panel study (de Vaus, 2001). The essential distinction between the two approaches lies in whether the same cases (individuals) are observed or surveyed over time, or which levels of changes are surveyed. In a panel study, the data is collected by the repeated survey of targeted individuals, which are used to detect both aggregate and individual changes. In contrast, a trend study often collects the data from comparable groups or regions over time but not from the same cases or individuals inside these groups or regions. A trend study is one specific case of panel study, since trend study can only reflect changes at the aggregated level, but not at the individual level. Thus, the method of trend study is specific to a research which is difficult to be designed as a panel study (for example, due to the unavailability of data collection), while still pursues the tendency of the effects of aggregate-level intervention/events/treatments.

The study in this paper belongs to a trend study, because the data was collected from different individuals nested within the same NPAs which are evolving over time (from 1997 to 2006). In fact, a study on land use-travel connection is hard to design as a panel study, since the panel data collection would be very costly, as confronted by many sociological studies (Rindfleisch et al. 2008), and old residents would move out of or new residents would move into one particular neighborhood. In addition, a costly panel design may be not necessary for a study on land use-travel connection, because the land use-travel study mainly aims at evaluating the efficiency of specific land use policies or planning strategy, which are commonly performed at neighborhood or city level rather than at household or individual level.

A large number of statistical models have been developed to analyze panel data, such as repeated-measure Univariate Analysis of Variance (UANOVA), and Multivariate Analysis of Variance (MANOVA), multilevel models which is also known as Hierarchical Linear Models (HLM, Raudenbush & Bryk, 2002), random coefficient models (Hsiao and Pesaran, 2008), and mixed-effect models (Neter et al., 1996). However, relatively fewer modeling frameworks have been developed to conduct a trend study.

Consequently, the paper focuses on three approaches for a trend study using a longitudinal and multilevel data set that collected neighborhood-level land use data and individual-level travel data in 1997 and 2006 in Austin, Texas. These three approaches applied here include a comparison approach with two cross-sectional multilevel models using single-year data, a pooling approach by building one multilevel model with two-year data, and a longitudinal multilevel model following the modeling framework in Randenbush (1989), who developed a three-level longitudinal multilevel model to estimate school effects in education.

This paper is organized as follows: Section 2 describes data and methodologies, focusing on the modeling specifications of the three specific approaches; Section 3 present major empirical modeling results; Section 4 concludes the paper.

2 Data and Methodology

2.1 Research Areas and Data

The major data source for this study comes from 1997 and 2006 Household Travel Survey in Austin, TX. Both surveys were conducted by the Capital Area Metropolitan Planning Organization (CAMPO), which coordinates regional transportation planning with five counties and cities in the capital area of Texas. These two surveys have similar survey process, by telephone, and similar contents for investigation, which mainly record household information, individual characteristics, and a 24-hour travel diary of surveyed residents. Household information includes a household's location (longitude and latitude), size, car ownership, housing situation, income, and other factors affecting residential choice, etc. Individual characteristics include each household member's gender, age, driving license status, employment status, ethnicity, and so on. The most important part of a household travel survey is the travel diary, recording the trip-related information such as each trip's origin and destination locations, origin and end timing, travel modes, travel purposes, and so on.

1997

2006

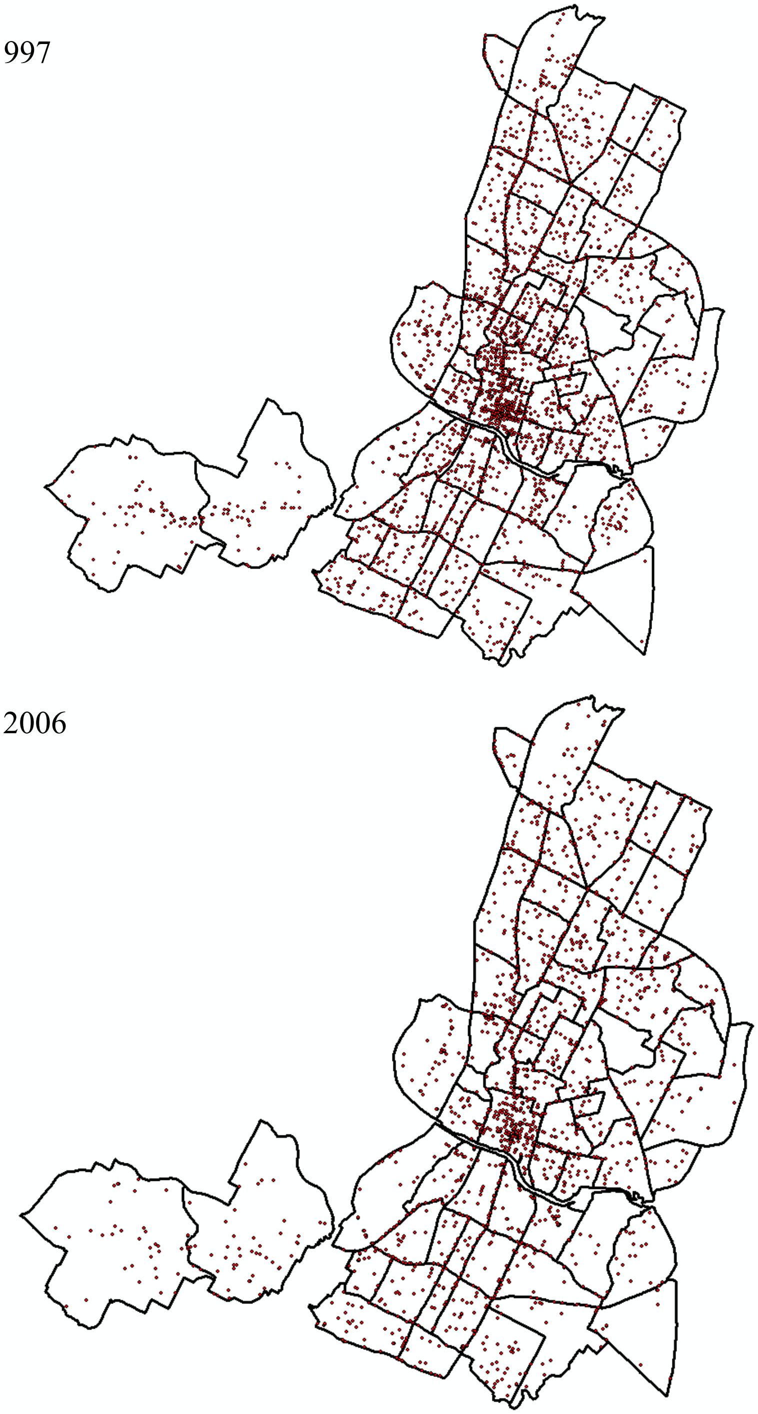

Figure 2.1 Research Neighborhoods and Surveyed Households (dots) in 1997 and 2006 in Austin, TX

This study only focuses on those trips originated at the neighborhood planning areas (NPA), which were divided by their historical and political coherence. In Austin, there are about 65 combined neighborhood planning areas, where the citizens are involved for participating in the planning process and decide how their neighborhoods will move into the future. There are two reasons why the NPA is chosen as the basic spatial unit for multilevel analysis in this study. First, residents living in a NPA may be a group with closer social interaction. Second, land use policies and plans are mainly conducted at NPAs; thus, studying the NPA's land use effects on travel behaviors may provide planners more practical policy implication.

Figure 2.1 shows the spatial distribution of NPAs in Austin and all the trips whose origin locations are inside these NPAs. The final dataset in this thesis includes 6849 trips surveyed in 1997 and 3231 trips surveyed in 2006. According to the spatial distribution of the sample set, 60 of 65 NPAs have trip origins located in.

2.2 Variables and Description

Since this study focuses on the determinants to travel mode choice between driving and non-driving (including public transit, walking, bicycling), the dependent variable, named as "Non-driving" is set as 1 when a trip is mainly finished by non-driving mode and as 0 by an automobile mode. From 1997 to 2006, the share of non-driving trips sharply drops from 16% to 9%, a net decrease of about 7%. This tendency shows increasing auto dependence in the studied NPAs at Austin; a growing number of trips rely on auto driving rather than public transit and walking.

Based on the data availability in both 1997 and 2006 travel surveys, this study only includes two categories of explanatory variables: individual-based socioeconomic attributes and neighborhood-based land use attributes. After eliminating several socioeconomic variables with low response rate in the survey, this study only involves variables such as time of trip, household size, number of cars, household income index, gender, race and age as socioeconomic attributes. The household income index is coded as from 1 to 15, with 1 representing $0-$5,000, 15 representing $150,000 or more, and an increment of $5,000.Obviously, a larger number of the index demonstrates a higher level of income. From 1997 to 2006, the mean time of trips significantly decreases from 16.58 minutes to 11.57 minutes, a net reduction of 5 minutes. In addition, household size, household car ownership, as well as household income display mild increase on the average. Demographically, the percentage of females shows a slight rise, and the same conditions remain for age distributions, while the proportion of whites drops about 10% percent.

On the other side, land use attributes contain population density (person per acre), household median income, and land use mix entropy. The population density kept stable between 1997 and 2006. However, remarkably, a more than $10,000 increase in household median income occurred during the decade. Land use mix entropy is calculated by the formula, $\text{Entropy} = -\sum_j [P_j * \ln(P_j))]/\ln(J)$, where Pj is the proportion of developed land in the jth land-use type and J is the number of land use categories considered. In this study, J = 6: residential, commercial, office, industrial, civic and open space. A larger value of this index indicates a higher level of mixed land-use pattern in the neighborhood. There was only about 0.02 increase in the land use mix entropy from 1997 to 2006.

Table 2.1 Descriptive analysis of variables in 1997

Variables	Mean	Std. Dev.	Min	Max
Dependent Variables				
Non-Driving (100%)	0.162	0.369	0	1
Socioeconomic Attributes				
Time of Trip (minutes)	16.582	30.998	1	924
Household Size	2.889	1.456	1	10
Number of cars	1.779	0.930	0	7
Household Income (1-15)	7.183	3.648	1	15
Female (100%)	0.524	0.499	0	1
White (100%)	0.650	0.477	0	1
Age	35.573	17.335	1	92
Land Use Attributes				
Population Density (persons per acre)	7.544	4.288	0.042	23.432
Household Median Income ($)	29,292	12,148	8,601	54,985
Land Use Mix Entropy	0.716	0.105	0.410	0.906

Table 2.2 Descriptive analysis of variables in 2006

Variables	**Mean**	**Std. Dev.**	**Min**	**Max**
Non-Driving (100%)	0.091	0.288	0	1
Socioeconomic attributes				
Time of Trip (minutes)	11.574	8.548	1	90
Household Size	3.318	1.737	1	10
Number of cars	1.877	0.792	0	6
Household Income (1-15)	7.901	3.726	1	15
Female (100%)	0.549	0.498	0	1
White (100%)	0.543	0.498	0	1
Age	39.435	21.139	0	94
Land use attributes				
Population Density (persons per acre)	7.409	3.696	0.759	21.242
Household Median Income ($)	41,391	16,073	12,758	80,789
Land Use Mix Entropy	0.735	0.124	0.255	0.914

2.3 Methodology: Longitudinal Multilevel Model (LMM)

Both the statistical methodologies of longitudinal and multilevel designs have recently received increasing attention in a variety of realms, such as education, psychology, sociology, and econometrics (Raudenbush, 1989; Kwok, 2008). The benefits of longitudinal data and analysis, generally comparing to cross-sectional data and analysis, are widely acknowledged in academics, including (1) longitudinal designs entail exploring the developments or changes over time in the outcomes of the target population at both the group and the individual level (de Vaus, 2001); (2) longitudinal designs "enhance the validity of causal inferences in nonexperimental research by providing a basis for assessing the direction of causation between two variables and by enabling some control over selection effects" (Raudenbush, 1989). Meanwhile, incorporating a multilevel analysis into statistical analysis enables the researchers to model the dataset with multilevel structure and "avoid a variety of errors of statistical inference including heightened probabilities of Type I errors, aggregation bias, and undetected heterogeneity of regression" (Raudenbush, 1989).

The dataset used in this paper have both longitudinal and multilevel elements: both the dependent and independent variables were measured over two years; one part of the independent variables (i.e., socio-economic attributes) is measured at the individual level, while the other part (i.e., land use characteristics) is measured at the neighborhood level. There may be three types of method/strategy to conduct a trend study, including a comparison model based on multi-wave cross-sectional analyses, a pooling model bringing in time-variant effects, and a longitudinal multilevel model. The following sections will provide statistical specifications of the three types of models for trend studies applied in this paper.

2.3.1 A Comparison Model based on Two-year Cross-Sectional Analyses

A simple methodology to cope with two-year data is to build two separate cross-sectional models, and then compare the changes of coefficients of each variable. Considering a single year's data structure is multilevel, consisting of individual and neighborhood levels, an appropriate model for the cross-sectional analysis on the effects of travel mode choice is a two-level multilevel model. The dependent variable here is a discrete variable measuring whether a trip uses driving or non-driving mode. Since it is not a continuous variable, a multilevel Logit model is necessary here.

In specific, the discrete outcomes of travel mode choice traditionally fit the discrete choice model under the assumption of random utility maximization (RUM). Consider an individual *i (i = 1,2, …, I)* living in neighborhood *j (j = 1,2,…, J)* chooses an alternative of travel mode *m (m = 1,2,…, M)*, with the following utility function (Akiva and Lerman, 1985):

$$U_{ijm} = \alpha_{jm} + \boldsymbol{\beta}_j' \boldsymbol{z}_{ijm} + \boldsymbol{\gamma}_m' \boldsymbol{x}_{ij} + \epsilon_{ijm} \tag{1}$$

where α_{jm} is a scalar utility term for alternative *m* associated with the neighborhood *j*, $\boldsymbol{z}_{ijm}$ is an individual-specific covariate vector which varies over alternatives and may also vary over individual/neighborhoods whereas the vector $\boldsymbol{x}_{ij}$ varies over individuals but not alternatives. The coefficient vector ($\boldsymbol{\beta}_j$) of alternative-associated variables, cost and time here, is hypothesized to vary across neighborhoods. $\boldsymbol{\gamma_m}$ represents the coefficient vector of individual-associated variables, such as age, household income, and household size, etc., and only varies across alternatives. ϵ_{ijm} stands for an unobserved random term that captures some factors affecting the utility but not included. We assume ϵ_{ijm} to be independently and identically (IID) distributed.

Eq. (1) represents the individual-level variation of travel mode choice. Then, we allow the intercept term, α_{jm}, and the coefficient vectors, β_j, to vary across their living neighborhood:

$$\alpha_{jm} = \delta_m + \boldsymbol{\pi'_m w_j} + \theta_{jm} \tag{2}$$

In Eq. (2), δ_m is an alternative specific constant of the average effect of unobserved variables on the utilities associated with travel mode m. w_j represents specific land-use variable, including density, design, accessibility and diversity, etc. π_m is a corresponding coefficient vector with respect to mode m. $\theta_{jm} \sim N(0, \sigma_m^2)$ is a random term that represents unobserved idiosyncratic differences across neighborhoods, and is also assumed to be IID distributed.

Combining formulations (1) and (2), the integrated equation will be:

$$U_{ijm} = (\delta_m + \boldsymbol{\pi'_m w_j} + \boldsymbol{\beta'_j z_{ijm}} + \boldsymbol{\gamma'_m x_{ij}}) + (\theta_{jm} + \epsilon_{ijm}) \tag{3}$$

The segment$(\delta_m + \boldsymbol{\pi'_m w_j} + \boldsymbol{\beta'_j z_{ijm}} + \boldsymbol{\gamma'_m x_{ij}})$ in Eq. 3 contains the fixed coefficients whereas the segment $(\theta_{jm} + \epsilon_{ijm})$ contains the random error terms. Choice model involving multilevel structure makes the Independence from Irrelevant Alternatives (IIA) property of multinomial Logit model (MNL) fail, as there exists dependence between individuals within neighborhoods induced by unobserved heterogeneity between neighborhoods (Bhat, 2000).

Letting the error terms θ_{jm} $(m = 2, \ldots, M)$ be conditioned, the probability of choice of mode *m* for individual *i* nested within neighborhood *j* can be written in the traditional MNL form:

$$P_{ijm} | \left(\theta_{j2}, \ldots, \theta_{jM},\ \varphi_2, \ldots, \varphi_K\right) = \frac{exp\ (\delta_m + \pi_m' w_j + \beta_j' z_{ijm} + \gamma_m' x_{ij} + \theta_{jm})}{\sum_n^M exp\ (\delta_n + \pi_n' w_j + \beta_j' z_{ijn} + \gamma_n' x_{ij} + \theta_{jn})} \quad (4)$$

Since the unconditional likelihood function for the multilevel choice model do not have closed form solutions, the maximization of the likelihood requires an integral approximation. In this study, I adopt adaptive Gauss-Hermite quadrature to integrate out the latent variables and obtain the marginal log-likelihood under program gllamm in STATA (Rabe-Hesketh et al. 2004). Applying an adaptive quadrature technique is supposed to provide a promising improvement of the accuracy of approximate estimation methods, rather than marginal quasi-likelihood, penalized quasi-likelihood and Gaussian quadrature (Rabe-Hesketh et al. 2002). However, adaptive quadrature estimation in program gllamm could be very slow, particularly if the models include many random effects (Rabe-Hesketh et al. 2002). We should first statistically test which coefficients in the Level-1 model significantly vary across neighborhoods, and which random effects can be combined, in order to balance between accuracy and efficiency.

After estimating the multilevel Logit models based on two years' data, we are able to compare both the coefficients of fixed and random effects between the two years.

From this comparison, one may address the differences of the neighborhood-based land use effects or individual-based socioeconomic effects between 1997 and 2006.

2.3.2 A Pooling Model with Time-Variant Effects

Similar to the comparison model previously addressed, a pooling model is a two-level multilevel Logit model as well. The difference lies in that in a pooling model, all two-year data come into one single model, assuming the variance of the residual to be the same in 1997 and 2006. A major advance of this pooling model is increased precision in estimation, since pooling two years' data together brings an increase in the number of observations and consequentially a larger sample size.

In order to estimate the time-variant effects, the pooling model can involve several time-related variables, for example, a dummy variable measuring a specific year (e.g., the variable Year_1997 represents 1 for 1997 and 0 for 2006), and the cross terms calculated by the year-related dummy variable times each land use variable.

2.3.3 A Longitudinal Multilevel Model

Both the comparison model and the pooling model are not sufficient to accurately estimate the dynamic feature of land use effects on travel mode choice. For example, the comparison model with two separated cross-sectional models can tell researchers the difference of land use effects on travel mode choice between the two years; however, it cannot allow researchers to distinguish between stable (time-invariant) and unstable (time-variant) components underneath the land use effects. The pooling model enables an estimation of stable and unstable land use effects after involving relevant time-related variables; however, it may lead to underestimated standard errors and t-statistics and overstates the precision gains, since it does not allow the variation of the residual to change over time. Therefore, a more accurate longitudinal model is needed to analyze the trends of land use effects on non-driving choice, i.e., a longitudinal multilevel model.

The structure of the land use-travel data used in this paper is similar to the data structure in several educational research about trends of school effects on student achievement. Many early studies have established models to capture the dynamics of school effects (Marco 1974; Rowan and Denk 1982; McPherson and Willms 1987; Randenbush 1989); variation among school effects includes both stable and unstable components, and it is important to distinguish between them (Willms and Randenbush 1989). In order to use the model originated from the realm of education, we can make an analogy that regards "neighborhood effects" (from land use) as "school effects", and "individual residents" as "individual students". In this section, we follow the longitudinal multilevel model developed by Raudenbush (1989), and extend it to a Logit outcome.

Specifically, the longitudinal multilevel model built in this thesis involves three levels of equations. In the first level (between-individual model), a separate regression of outcomes is conducted on individual-level socioeconomic variables within each neighborhood at each point of time. The utility function of each individual (or each trip) is thus become the formation as follows:

$$U_{ijt}^{m} = \beta_{jt0} + \sum_{k=1}^{K} \beta_{jtk} \mathrm{x}_{\mathrm{jtk}} + \epsilon_{ijt} \tag{5}$$

for individual i (i = *1,2, …, I*) living in the neighborhood j (j = *1,2,…, J*) chooses an alternative of travel mode m (m = driving, non-driving) at occasion t (t = 1997, 2006). Notice that at each occasion t, this model turns to be identical to the multilevel Logit model as described in Eq. (1). Since there are totally 60 neighborhoods observed on two occasions, Eq. (6) would specify 120 separate regressions. The β_{jt0} are estimates of the neighborhood amenities (that benefit for non-driving choice) for each neighborhood j at location t, after adjusting for covariates in the model.

The second level (between-neighborhood model) in the model is a regression of outcomes on neighborhood-level land use variables at each point in time, which allows for the individual-level effects varying between neighborhoods. Here we assumed that only the constant item in Level-1 is allowed to change between neighborhoods, as displayed in Eq.(6).

$$\beta_{jt0} = \pi_{t0} + \pi_{t1} w_{t1} + \pi_{t2} w_{t2} + \pi_{t3} w_{t3} + \theta_{jt0} \tag{6}$$

The effects of socioeconomic variables are assumed to have no between-neighborhood variance. w_{tk} (k = 1, 2, 3) represents the land use variable k at occasion t, including population density, household median income, and land use mix entropy. π_{tk} (k = 1, 2, 3) are related coefficients of the three land use variables. The model in this level allows one to specify how much of the instability can be explained by changes in neighborhood-based land use policies and contexts.

As to the third level (between-occasion model), the estimates of the constant in the neighborhood effect (δ_t) and the coefficients of the relevant land use variables (π_{tk}) are regressed on the time point of the t^{th} observation for each neighborhood.

$$\pi_{t0} = \gamma_{00} + \gamma_{10}\rho_t + \varepsilon_{t0} \tag{7}$$

$$\pi_{tk} = \gamma_{0k} + \gamma_{1k}\rho_t,\ k=1,\ 2,\ 3 \tag{8}$$

where, ρ_t is a time dummy variable, which is assigned 1 for 1997 samples and 0 for 2006 samples. γ_{0k} and γ_{1k} (k = 0, 1, 2, 3) represent corresponding coefficients. Eq. (7) indicates that the average constant of neighborhood effects vary across time points from 1997 to 2006. ε_{t0} represent random year-to-year fluctuations in the neighborhood level. Eq. (7) is developed by assuming that all the slopes of the land use variables will be changed over years.

Combining equations (5)-(8) yields the simple model as follows:

$$\begin{aligned}
U_{ijt}^{m} &= \gamma_{00} && \text{(grant mean)} \\
&+\gamma_{10}\rho_t && \text{(main effect of time)} \\
&+\beta_1(Time\ of\ trip)_{ijt} + \beta_2(HH\ Size)_{ijt} + \beta_3(Car\ Ownership)_{ijt} + \\
&\quad \beta_4(HIcome)_{ijt} + \beta_5(Female)_{ijt} + \beta_5(White)_{ijt} + \beta_5 Age_{ijt} \\
& && \text{(control for individual-level attributes)} \\
&+\gamma_{01}(Density)_{jt} + \gamma_{02}(MIncome)_{jt} + \gamma_{03}(Mix)_{jt} + \theta_{jt0} \\
& && \text{(Stable component of land use effects)} \\
&+\rho_t\left[\gamma_{11}(Density)_{jt} + \gamma_{12}(MIncome)_{jt} + \gamma_{13}(Mix)_{jt}\right] + \varepsilon_{t0} \\
& && \text{(unstable component of land use effects)} \\
&+\epsilon_{ijt} && \text{(individual error)}
\end{aligned} \tag{9}$$

3 Trends of Land Use Effects on Travel: Comparing Three Models

In this chapter, three models described above (comparison model, pooling model, and longitudinal multilevel model) would be estimated and compared using the realistic data to study the trends of land use effects on patterns of travel behavior in Austin. Basically, we will compare the goodness-of-fit statistics of the three models, and explore which types of model can provide a better understanding on the trends of land use effects on the choice of non-driving travel mode, particularly referring to public transit, bicycling, and walking.

3.1 The Comparison Approach: Two-year Cross-sectional Multilevel Logit Model

The comparison model consists of two separate cross-sectional multilevel Logit models. As is shown in Table 3.1, by comparing the two modeling results, we can explore the distinctions in the determinants of travel mode choice between 1997 and 2006. The Log-likelihood value in the model representing 1997 is -1411.653, while in the model of 2006 it is -818.151. The likelihood ratio tests prove that both models fit the data well.

In the model of 1997, the utility variance of the non-driving mode across neighborhood (in Level-2) is 0.294 with a standard error of 0.107, indicating that the multilevel model cannot collapse into a single-level regression model. Since the Level-1 variation of Logit structure can be assumed as $\pi^2/3$ (Hox, 2002), the between-neighborhood correlation turns to be 0.08, calculated by 0.294/(0.294 + $\pi^2/3$). This number indicates that more than 8% of the variation of the log-odds of non-driving mode probability occurs between neighborhoods contexts, and that nearly 92% of the variation is explained by the individual-level variables.

In the model of 2006, the multilevel model is also statistically needed, since the utility variance of the non-driving mode across neighborhood is 0.693 with a standard error of 0.232. The between-neighborhood variation accounts for about 17%, calculated by 0.693/(0.693 + $\pi^2/3$), of the total variation.

First, the constants are both insignificant in 1997 and 2006 after controlling for the effects of socioeconomic and land use variables, but become more negative in 2006. Subjects in 2006 might have a higher probability of driving for transportation rather than non-driving modes in their daily trips, comparing to 1997. This finding appears to indicate an increasing tendency of auto dependence in Austin.

Second, similar results are found in the two year analyses about the effects of Level-1 socioeconomic variables on travel mode choice. For example, longer time of trips significantly reduce the possibility of using non-driving modes. Larger household size may generate more non-driving trips, but the coefficients of both years are insignificantly. More numbers of cars significantly encourage people to drive more, or vice versa. The higher the household income, the lower the probability of non-driving; but this relationship is only significant in 1997's data. In addition, both the female and the elder significantly prefer to drive rather than use non-driving modes, while the race difference on travel mode choice seems statistically insignificant.

Table 3.1 Two Multilevel Logit Models with 1997 and 2006 Data

Fixed Effects	**Model 1: 1997**			**Model 2: 2006**		
	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z
Level-1						
Constant	-1.207	1.058	0.254	-2.102	1.403	0.134
Time of Trip (minutes)	-0.021	0.005	0.000	-0.001	0.008	0.902
Household Size	0.056	0.039	0.151	0.020	0.044	0.649
Number of cars	-0.477	0.072	0.000	-0.720	0.106	0.000
Household Median Income (1-15)	-0.047	0.018	0.007	-0.024	0.023	0.306
Female (100%)	-0.380	0.102	0.000	-0.326	0.137	0.017
White (100%)	0.223	0.133	0.094	-0.180	0.168	0.284
Age	-0.033	0.004	0.000	-0.025	0.004	0.000
Level-2						
Population Density (persons per acre)	0.085	0.028	0.003	-0.011	0.045	0.813
Average Household Income ($)	0.000	0.000	0.088	0.000	0.000	0.174
Land Use Mix Entropy	1.097	1.062	0.301	3.460	1.414	0.014
Random Effects						
var(Level-2):	0.294	0.107		0.694	0.232	
Log-Likelihood	-1411.653			-818.151		

Third, in the Level-2 model, the land use effects on travel mode choice tend to vary largely between 1997 and 2006. The variable of population density of neighborhoods has a significant effect on individual's non-driving choice in 1997, but insignificant in 2006. The effects of the income level of neighborhoods are not significant at the 0.05 level for both the 1997 and 2006 models. People living in more mixed-use neighborhoods have a higher probability to travel by non-driving modes in both models; but only the effect of land use mixture in 2006 is significant. These findings suggest that the roles of different land use policies, such as densification and mixed use development, in encouraging non-driving travel may vary across time points. Thus, it is important to estimate the time-invariant and time-variant components of such land use effects. However, comparing two cross-sectional models cannot successfully measure these changes as desired.

3.2 The Pooling Multilevel Model

The pooling model introduces two years data into one model by involving time-related variables. Table 4.2 shows the estimated results of two pooling models: one model only involves a time dummy variable (Model 3), while the other model not only has a time dummy variable but also three cross terms multiplying the dummy variable of time by the three variables of land use feature (Model 4).

In model 3, the estimated effects of individual-level socioeconomic variables on non-driving choice are similar to those in the cross-sectional models (Model 1 & Model 2). The explanatory variables such as time of trip, number of cars, household income and age are proved to be significant determinants to the travel mode choice between driving and non-driving modes, while factors such as household size, gender, and race are not statistically significant determinants. The pooling modeling results also shows that both the neighborhood-based population density and land use mix entropy exert a significant impact on the mode choice of trips. People living in a neighborhood planning area with a higher population density and a more mixed land use pattern hold a higher possibility to make use of non-driving travel mode. Nevertheless, the effect of neighborhood household income level appears to be insignificant.

The coefficient of the year-dummy variable (Year_1997) is estimated at -0.344 with a standard error of 0.125 and a p-value of 0.006. This finding indicates that there is a significant difference of non-driving mode choice between the two target occasions, year1997 and 2006. This finding also suggests that a pooling multilevel model may provide more information for comparing the time-variant determinants to non-driving choice that the comparison model, since the pooling model is able to statistically test the time-variant assumptions. In Model 3, it is found that the surveyed trips in 2006 significantly rely more on non-driving modes than those in 1997, after controlling for socioeconomic and land use variables. This finding, however, tends to contradict the finding from the previous descriptive analysis and comparison model, which tend to detect an increasing auto dependence and a decreasing non-driving use from 1997 to 2006.

In the meantime, Model 4 presents a more comprehensive estimation to figure out the change of different land use effects on non-driving choice. From Model 3 to Model 4, the Log-likelihood values increase from -2254.866 to -2247.46. A likelihood ratio test can provide evidence that Model 4 is considerably better than Model 3 (p-value = 0.05). Here we can interpret both the time-invariant and time-variant land use effects. For example, population density has no significant stable effects on non-driving choice (p-value = 0.833), but shows significant time-variant effects (p-value = 0.001). This contrast indicates that the effects of population density on travel mode choice may meaningfully vary across time.

In contrast, the time-invariant effect of land use mixture is significant (p-value = 0.001), while the corresponding time-variant effect displays no significance (p-value = 0.269). This piece of result implies that the effects of mixed land use policies are relatively stable across years. Analogical results are reached on the effects of neighborhood's median household income. However, the coefficient of median household income becomes significant at the 0.05 level, which appears to contradict the finding in the comparison model: in both the two cross-sectional models for year 1997 and 2006, no significant links between median household income and non-driving probability are detected. After controlling for the time-variant effects of all the three involved land use variables, the coefficient of the time-dummy variables converts to be strongly insignificant.

In summary, by means of the pooling multilevel Logit model, we can distinguish not only the time-variant but also the time-invariant land use effects on non-driving mode choice, and also address their statistical significance. However, some findings appear to contradict a previous statement in the descriptive analyses as well as the cross-sectional modeling analyses, such as the effects of year-dummy variable in Model 3 and the effects of median household income in Model 4. This contradiction may result from the strong assumption underlying the pooling model that the variation of the residual is fixed over time. However, this assumption is probably open for challenge because the variation of the residual on people's travel mode choice may largely change over the decade from 1997 to 2006. Thus, to pursue a more accurate estimation, the solution should rely on a model allowing for time-variant variances and/or covariance, i.e., a longitudinal multilevel model.

Table 3.2 Pooling multilevel model with time-varying variables

Fixed Effects	Model 3			Model 4		
	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z
Level-1						
Constant	-2.734	0.840	0.001	-1.992	1.044	0.056
Time of Trip (minutes)	-0.017	0.004	0.000	-0.017	0.004	0.000
Household Size	0.026	0.028	0.349	0.035	0.028	0.215
Number of cars	-0.553	0.059	0.000	-0.561	0.059	0.000
Household Income (1-15)	-0.040	0.014	0.003	-0.039	0.014	0.004
Female (100%)	-0.356	0.081	0.000	-0.358	0.081	0.000
White (100%)	0.068	0.102	0.508	0.068	0.103	0.509
Age	-0.029	0.003	0.000	-0.029	0.003	0.000
Level-2						
Population Density (persons per acre)	0.077	0.025	0.002	0.007	0.031	0.833
Average Household Income ($)	0.000	0.000	0.175	0.000	0.000	0.022
Land Use Mix Entropy	3.339	0.862	0.000	3.481	1.059	0.001
Time-Related Variables						
Year_1997	-0.344	0.125	0.006	-0.197	1.069	0.854
Population Density *Year_1997				0.085	0.025	0.001
Household Median Income *Year_1997				0.000	0.000	0.999
Land Use Mix Entropy *Year_1997				-1.229	1.112	0.269
Random Effects						
var(Level-2):	0.353	0.114		0.325	0.104	
Log-Likelihood	-2254.866			-2247.462		

3.3 Longitudinal Multilevel Model

The longitudinal multilevel model illuminated from Raudenbush (1989) can combine both the benefits of the comparison model and the pooling model, as it allows one to estimate the time-invariant and time-variant effects and variations. As described in last section, we establish two comparable longitudinal multilevel models: one model only involves a time-dummy variable (Model 5), while the other model includes not only the time-dummy variable but also three cross terms (Model 6).

Both the log-likelihood ratio tests for comparing in pair Model 5 (log-likelihood=-2244.535) and Model 3 as well as Model 6 (log-likelihood=-2240.795) and Model 4 demonstrate that Model 5 and Model 6 performs better in fitting the data stemming from that they internalize the change of variations over time in the model design.

The socioeconomic effects in the Level-1 model of Model 5 are very similar to those in Model 3. The significance keeps consistent for each variable in the two models, but all the related standard errors in Model 5 become slightly larger than those in Model 3; In the Level-2 model, the coefficients of the three land use variables don't change a lot as those in Model 3, while the p-values of these variables delightedly decrease to be all significant in Model 5. By Model 5, people living in a neighborhood with higher density, higher income, and more mixed land use pattern are more likely to use non-driving mode for their daily trips. The coefficients of density and land use mix are both significant at the 0.01 level, whereas that of neighborhood median income is insignificant at the 0.05 level.

The coefficient of the time dummy variable in Model 5 is -0.376 with a p-value of

0.077, showing that there is an insignificant (at the 0.05 level) main effect of time. This finding appears to be more reasonable than the pooling model, by being more consistent with those findings in the descriptive analysis and the comparison model. The longitudinal multilevel model can provide more accurate results as it would not lead to underestimated standard errors and t-statistics and thus would not overstate the precision gains, since it allow for the variation of the residual to change over time (in the Level-3).

By comparing Model 6 with Model 4, we can again experience the statistical advantages of a longitudinal multilevel model. The findings in Model 6 of the land use effects on non-driving mode choice tend to be more accurate. For example, both the time-variant and time-invariant effects of neighborhood median income are insignificant at the 0.05 level. This finding appears to be more accurate than that in the pooling model, which on the contrary reveals that the median income has a significant effect on non-driving mode choice. The effects of population and land use mix are similar to those in Model 4; but their standard errors become larger.

In summary, the longitudinal multilevel model generally performs better than both the comparison model and the pooling model via enabling researchers to estimate the time-variant and time-invariant fixed effects, as well as the time-variant random effects. This comprehensive model is more productive in analyzing the longitudinal land use effects on travel behaviors. By analyzing Model 5 and Model 6, we suggest that the land use policy of increasing the mixture of land use of a region is a robust incentive for encouraging less driving and more non-driving trips, while policies that improve land use density may have lead to an increasing use of non-driving modes only in 1997, but not in the recent year of 2006. We may use a dynamic perspective to explain this phenomenon. In 1997, most of the Austin neighborhoods were still hanging in a low-density status due to the urban sprawling development. Public transit facilities were therefore rarely imputed to the low population density, and thus residents had to depend on driving to a large extent. Under this condition, there was space to enhance the population density in many of neighborhoods located in Austin, and hence a densification policy promised to be remarkably effective in reducing driving behavior. However, in 2006, population density in neighborhoods had already increased largely and even approached the peak value due to many densification policies conducted in Austin, such as Compact Development and Smart Growth. In this case, continuing to raise density may not be an effective policy for encouraging less driving. This finding seems consistent with some recent studies in urban planning field, which reveal many negative effects of over-densification policy (Chatman, 2008).

Table 3.3 Longitudinal multilevel model

Fixed Effects	Model 5			Model 6		
	Coef.	Std. Err.	P>z	Coef.	Std. Err.	P>z
Level-1						
Constant	-1.966	0.901	0.029	-2.049	1.235	0.097
Time of Trip (minutes)	-0.017	0.004	0.000	-0.017	0.004	0.000
Household Size	0.037	0.029	0.200	0.040	0.029	0.166
Number of cars	-0.541	0.059	0.000	-0.543	0.059	0.000
Household Income (1-15)	-0.040	0.014	0.004	-0.040	0.014	0.004
Female (100%)	-0.354	0.082	0.000	-0.353	0.082	0.000
White (100%)	0.071	0.103	0.491	0.073	0.103	0.482
Age	-0.029	0.003	0.000	-0.029	0.003	0.000
Level_2						
Population Density	0.057	0.022	0.010	-0.018	0.039	0.644
Average Household Income	0.000	0.000	0.054	0.000	0.000	0.090
Land Use Mix Entropy	2.657	0.898	0.003	3.551	1.252	0.005
Level_3						
y1997	-0.376	0.213	0.077	0.783	1.731	0.651
Population Density *Year_1997				0.114	0.051	0.026
Household Median Income*Year_1997				0.000	0.000	0.648
Land Use Mix Entropy *Year_1997				-2.505	1.775	0.158
Random Effects						
var(Level-2):	0.553	0.135		0.484	0.118	
var(Level-3):	0.001	0.011		0.000	0.006	
Log-Likelihood	-2244.535			-2240.795		

4 Conclusion

Land use and transportation planners have conducted numerous studies about land use effects on travel mode choice, since less driving and more non-driving behaviors such as riding a public transit, bicycling, and walking, are widely regarded as more sustainable travel behaviors that provide support for solving many environmental, energy and social equity issues. However, most of the previous studies rely on the methodologies developed by cross-sectional data; less research in the land use-travel connection explore the statistical techniques for longitudinal design and analysis. This thesis thus attempts to establish three distinct modeling frameworks for longitudinal study, using empirical data from two years (1997 and 2006) in the city of Austin, and then compares them to determine which one is better for analyzing longitudinal and multilevel datasets.

In the methodology section, this thesis summarizes two types of longitudinal study: a panel study and a trend study. We found that the longitudinal land use-travel study often belongs to a trend study, which collects data from comparable regions (e.g., neighborhoods, census tract, and traffic analysis zones) over time but not from the same cohort of individuals inside these regions, considering that the repeated individual panel survey is difficult to conduct in the field of city planning. Also, when planners focus on land use policies, which are usually carried out at the neighborhood or city level rather than the individual level, they can only rely on a trend study, which will be sufficient and convenient for them to understand the changes at the aggregated level.

Three longitudinal modeling design strategies in this thesis are compared, including a comparison two-level multilevel model by building two separate cross-sectional multilevel Logit models, a pooling model by combining two years' data into one multilevel Logit model and adding several time-variant effects, and a longitudinal multilevel model by involving an additional level (Level-3) to measure the effect between the two time points. The empirical modeling results show that the comparison model allow the researchers to detect the difference of land use effects on non-driving mode choice in 1997 and 2006, but cannot distinguish the time-invariant and time-variant effects. The pooling model performs better to estimate the time-invariant and time-variant land use effects, but may lead to underestimated standard errors and t-statistics and may overstate the precision gains, since the pooling model does not allow the variation of the residual to change over time. The longitudinal multilevel model turns to perform the best in the three models including the comparison model and the pooling model, since it allows researchers to estimate the time-variant effects and the fixed effects that generally do not vary with time, as well as the time-variant random effects.

References

Akiva, M. E. B., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to predict travel demand* (Vol. 9). Cambridge, MA: The MIT press.

Badoe, D. & Miller, E. (2000). Transportation-Land Use Interaction: Empirical Findings in North America, and Their Implications for Modeling. *Transportation Research D, 5*, 235-263.

Bhat, C. (2000). A Multi-level Cross-classified Model for Discrete Response Variables. *Transportation Research Part B*, *34*, 567-582.

Bhat, C., Carini, J. & Misra, R. (1999). On Modeling the Generation and Organization of Household Activity Stops. *Transportation Research Record, 1676*, 153-162.

Bhat, C., & Gossen, R. (2003). A Mixed Multinomial Logit Model Analysis of Weekend Recreational Episode Type Choice. Paper presented at the 2003 Transportation Research Board annual meeting, Washington, D.C.

Boarnet, M.G. (2011). A Broader Context for Land Use and Travel Behavior, and a Research Agenda. *Journal of the American Planning Association, 77*(3), 197-213.

Cao, X., Mokhtarian, P. & Handy, S. (2009). The Relationship Between the Built Environment and Nonwork Travel: A Case Study of Northern California. *Trasnportation Research A*, *43*(5), 548-559.

Chatman, D. G. (2008). Deconstructing Development Density: Quality, Quantity and Price Effects on Household Non-work Travel. *Transportation Research Part A: Policy and Practice*, *42*(7), 1008-1030.

Chatman, D. (2009). Residential Choice, the Built Environment, and Nonwork Travel: Evidence Using New Data and Methods. *Environment and Planning A*, *41*(5), 1072-1089.

Chen, C. & Chen, J. (2009). What is Responsible for the Response Lag of a Significant Change in Discretionary Time Use: the Built Environment, Family and Social Obligations, Temporal Constraints, or a Psychological Delay Factor? *Transportation*, *36*(1), 27-46.

Crane, R. (1999). *The impacts of urban form on travel: A critical review*. Boston: Lincoln Institute of Land Policy.

Crane, R. & Crepeau, R. (1998). Does Neighborhood Design Influence Travel? A Behavioral Analysis of Travel Diary and GIS Data. *Transportation Research D*, *3*(4), 225-238.

De Vaus, D. (2001). *Research Design in Social Research*. London: Sage.

Ewing, R. & Cervero, R. (2001). Travel and the Built Environment: A Synthesis. *Transportation Research Record*, *1780*, 87-114.

Ewing, R. & Cervero, R. (2010). Travel and the Built Environment. *Journal of the American Planning Association*, *76*(3), 265-294.

Goulias, K. (1999). Longitudinal Analysis of Activity and Travel Pattern Dynamics Using Generalized Mixed Markov Latent Class Models. *Transportation Research B*, *33*(8), 535–557.

Goulias, K. (2002). Multilevel Analysis of Daily Time Use and Time Allocation to Activity Types Accounting for Complex Covariance Structures Using Correlated Random Effects. *Transportation, 29*, 31–48.

Greenwald, M. (2003). The Road Less Traveled: New Urbanist Inducements to Travel Mode Substitution for Nonwork Trips. *Journal of Planning Education and Research*, *23*(1), 39-57.

Greenwald, M. & Boarnet, M. (2001). Built Environment as Determinant of Walking Behavior: Analyzing Nonwork Pedestrian Travel in Portland, Oregon. *Transportation Research Record, 1780*, 33-42.

Handy, S. (1994). Regional Versus Local Accessibility: Implications for Nonwork Travel. *Transportation Research Record*, *1400*, 58-66.

Handy, S. & Clifton, K. (2001). Local Shopping as a Strategy for Reducing Automobile Travel. *Transportation*, *28*(4), 317-346.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* New Jersey: Lawrence Erlbaum Associates.

Hsiao, C., & Pesaran, M. H. (2008). Random Coefficient Models. In: Mátyás. L., & Sevestre, P. (eds). *The Econometrics of Panel Data.* Berlin Heidelberg: Springer.

Kwok, Y. K. (2008). *Mathematical models of financial derivatives*. Singapore: Springer.

Limanond, T., Niemeier, D.A. & Mokhtarian, P.L. (2005). Specification of a Tour-based Neighborhood Shopping Model. *Transportation*, *32*(2), 105-134.

Ma, J. & Goulias, K. (1997). An Analysis of Activity and Travel Patterns in the Puget Sound Transportation Panel. In: Ettema, D. F. & Timmermans, H.J.P. (eds). *Activity-Based Approaches to Travel Analysis*. The Netherlands: Pergamon.

Marco, G. L. (1974). A Comparison of Selected School Effectiveness Measures Based on Longitudinal Data. *Journal of Educational Measurement*, *11*, 225-234.

Matt, K. & Arentze, T. (2003). Variation of Activity Patterns with Features of the Spatial Context. Paper presented in the 82th Transportation Research Board annual meeting, Washington, D.C.

McPherson, A. F., & Willms, J. D. (1987). Equalisation and Improvement: Some Effects of Comprehensive Reorganisation in Scotland. *Sociology, 21*, 509-539.

Neter, J., Wasserman, W., & Kutner, M. H. (1996). *Applied linear statistical models*. Chicago: Irwin.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2002). Reliable Estimation of Generalized Linear Mixed Models Using Adaptive Quadrature. *The Stata Journal*, *2*, 1-21.

Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). Gllamm Manual. U.C. Berkeley Division of Biostatistics Working Paper Series.

Rajamani, J., Bhat, C., Handy, S., Knaap, G., and Song, Y. (2003). Assessing the Impact of Urban Form Measures in Nonwork Trip Mode Choice after Controlling for Demographic and Level-of-service Effects. Paper presented in the 78th Transportation Research Board annual meeting, January 12-16, 2003, Washington, D.C.

Raudenbush, S.W. (1989). The Analysis of Longitudinal, Multilevel data. *International Journal of Educational Research*, *13*(7), 721-740.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods. (Advanced Quantitative Techniques in the Social Sciences).* Thousands Oaks, CA: Sage.

Rowan, B., & Denk, C. E. (1982). *Modeling the Academic Performance of Schools Using Longitudinal Data: An Analysis of School Effectiveness Measures and School and Principal Effects on School-level Achievement.* San Francisco, CA: Instructional Management Program Publication, Far West Laboratories.

Rindfleisch, A., Malter, A.J., Ganesan, S. & Moorman, C. (2008). Cross-Sectional Versus Longitudinal Survey Research: Concepts, Findings, and Guidelines. *Journal of Marketing Research*, *45*(3), 261-279.

Schwanen, T. Dijst, M. and Dieleman, F. (2004) Policies for Urban Form and Their Impact on Travel: The Netherlands Experience. *Urban Studies*,*41*(3), 579–603.

Snellen, D., Borgers, A. and Timmermans, H. (2002). Urban Form, Road Network Type, and Mode Choice for Frequently Conducted Activities : a Multilevel Analysis Using Quasi-experimental Design data. *Environment and Planning A*, *34*(7), 1207-1220.

Willms, J. D., & Raudenbush, S. W. (1989). A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability. J*ournal of educational measurement*, *26*(3), 209-232.